DOCUMENT RESUME

ED 383 732 TM 023 172

AUTHOR Stecher, Brian
TITLE The Cost of Performance Assessment in Science: The RAND Perspective.
SPONS AGENCY National Science Foundation, Arlington, VA.
PUB DATE 18 Apr 95
CONTRACT NSF-MDR-9154406
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Classification; Comparative Analysis; *Costs; Educational Assessment; Elementary School Students; Grade 5; Grade 6; Inferences; Intermediate Grades; Multiple Choice Tests; Science Education; *Scoring; *Test Construction; *Testing
IDENTIFIERS Hands On Experience; Large Scale Programs; *Performance Based Evaluation; Rand Corporation

ABSTRACT

The resources necessary to create, administer, and score performance assessments in science were studied. RAND and the University of California, Santa Barbara (UCSB) designed performance tasks for science in grades five and six as part of a larger study of the feasibility of science performance assessment. Tasks were developed in pairs in task shells called inference and classification. A total of 2,200 students in both grades participated. Fifth graders completed the two UCSB tasks and sixth graders completed the four RAND tasks. Total costs for development, equipment preparation, task administration, and scoring were computed for each task. Results suggest that hands-on science testing is about 100 times as expensive as standardized multiple choice testing for a comparable amount of testing time and that it is five to six times as expensive as constructed response assessments in writing of comparable length. In addition, economies of scale associated with performance testing are limited. Task development costs drop dramatically as the number of students increases, but other costs do not. Implications for large-scale testing are considered. One figure and 11 tables present findings. (Contains 4 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 383 732

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRIAN STECHER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

THE COST OF PERFORMANCE ASSESSMENT IN SCIENCE: THE RAND PERSPECTIVE

Brian Stecher, RAND

Paper presented as part of the symposium "The Cost of Performance Assessment in Science" at the 1995 annual meeting of the National Council on Measurement in Education in San Francisco, California, April, 1995. This material is based upon work supported by the National Science Foundation under Grant No. MDR-9154406. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

BEST COPY AVAILABLE

THE COST OF PERFORMANCE ASSESSMENT IN SCIENCE: THE RAND PERSPECTIVE[1]

INTRODUCTION

Proponents of educational performance assessments claim that this form of testing has advantages over multiple choice tests both in terms of validity and consequences (Wiggens, 1989; Shavelson, et al., 1990). We will not debate these claims here. Instead, this paper focuses on the resources necessary to create, administer and score such assessments. It may be true that performance assessments embody more authentic aspects of domain performance and send better signals to teachers and students about desired instructional behaviors. However, policymakers need to weigh the potential benefits against the costs associated with performance testing when considering the use of such tests for large scale assessment.

Most educators are familiar with the costs of standardized multiple choice testing, and many base their expectations regarding the cost of alternative forms of assessment on their experience with commercial multiple choice tests. For example, the cost of the complete five hour CTBS battery (including reporting) is approximately $2.80 per student.[2] Using this basis, the 30 minute science subtest represents an investment of roughly $0.30 per student. Experience in the area of writing shows that the use of open-ended written responses raises the costs considerably. For example, the CTB writing assessment costs $4.80 per student for a single prompt (either one holistic score or one analytic score). One might assume that the costs of hands-on science testing, in which students complete open-ended responses, would be similar to the cost of constructed response writing assessment. Our experience suggest

---

[1]The author wishes to thank Stephen Klein and Randy Ross for their assistance and suggestions.

[2]Based on the use of reusable test books (three uses per booklet) and basic scoring and reporting services. Excluded are any accessory publications such as norm books, class management guides, technical reports or locator tests.

that this assumption is incorrect; the costs of hands-on science assessment in which students must construct, observe, measure, manipulate or otherwise interact with objects and equipment are considerably higher than the costs of written constructed-response assessments.

This paper is part of a symposium which will present initial estimates of the cost of developing, administering, and scoring hands-on performance assessment in science. Data from four different testing programs will help establish a reasonable estimate of the range of costs associated with this type of activity. It appears to us that science assessments are among the most costly performance tests to produce because of the added expense of equipment and materials, so these estimates may represent upper bounds for the cost of performance testing of similar scope in other subjects. However, these estimates do not include the resources necessary for analysis, reporting of results, or for other administrative functions associated with a large testing program, so they underestimate the total cost of a large-scale assessment system.

To facilitate comparisons between the four presentations in this symposium, all parties adopted common definitions and procedures, which are described in the next section. Following that, the chronology for the development, administration, and scoring of the RAND assessment is described. Information about the resources requirements, total costs, and costs per student of each task are presented next. Concluding remarks address the validity of assumptions underlying these analyses and the implications of the results for large-scale assessment.

DEFINITIONS AND PROCEDURES

For the purposes of this study, a <u>task</u> is defined to be a structured, hands-on performance event lasting approximately one class period (30 to 55 minutes), whose stimulus or solution involves manipulating scientific apparatus or materials. Students work on tasks independently and they produce written, tabular or pictorial responses. Student responses are assigned one or more scores by expert judges based on scoring criteria relevant to the task. This definition excludes

activities in which scores are assigned through direct observation of student activities, an approach which is rarely used and potentially far more expensive. For the purposes of this analysis, we also assume the tasks will produce student level scores. Reliable school-level scores can be achieved at lower cost by sampling students. In the RAND study we dealt only with individual tasks, but this definition of task does not exclude activities that have group elements, so long as students produce individual responses.

Although there are many steps in the assessment process, we agreed to describe the process in terms of four major activities, which are shown in Table 1 along with some common sub-activities.

Table 1

Stages of Assessment

Activity	Subactivity
Task Development	
	Domain Specification
	Initial Activity Development
	Equipment Prototype Development
	Score Guide Development
	Pilot Test and Revisions
	Field Test and Refinement
Equipment Preparation	
	Printed Material Preparation
	Equipment/Apparatus Acquisition
Task Administration	
	Material Distribution
	Administrator/Teacher Training
	Monitoring
	Material Collection
Scoring	
	Score Guide Refinement
	Reader Recruitment
	Reader Training and Calibration
	Scoring

We also agreed to delineate resources using the same general categories, which are shown in Table 2. For the purposes of these analyses, we omitted any estimate of the opportunity costs associated with the time teachers and students participated in field testing or in the actual assessment.

Table 2

Resource Categories

Major Categories	Sub-Categories
PERSONNEL	
	Senior Professional Staff
	Junior Professional Staff
	Research Assistants
	Consultants
	Teacher Consultants
TRAVEL	
EQUIPMENT	
CONTRIBUTED TIME	
	Teacher Time
	Student Time
CONTRACTED COSTS	
OTHER DIRECT COSTS	

To make the analyses comparable, we also agreed to the following assumption regarding the use of testing materials. When testing materials are disposable, e.g., solutions that must be combined, booklets in which students write responses, etc., the program should include one set of materials for each student. When materials are reusable, e.g., objects that are merely weighed or observed, test booklets that are not written upon, etc., the program should provide one set of materials for every three students. The three-to-one ratio is based on an estimate of the number of times a set of materials can be reused under standardized large-scale testing conditions, which usually require simultaneous testing in multiple classrooms. We discuss the validity of this assumption at the end of the paper.

RAND PERFORMANCE ASSESSMENT EXPERIMENT, SPRING 1993

These cost estimates are based on an experiment designed by RAND as part of a larger study of the feasibility of science performance assessment, funded by the National Science Foundation. The experiment was conducted during the spring of 1993 in Southern California in conjunction with the annual administration of the California Learning Assessment System (CLAS). RAND was assisted by researchers from the University of California, Santa Barbara (UCSB) and Stanford University, as well as by staff from CLAS.

Assessment Tasks

As part of that project, each institution developed a number of hands-on science assessment tasks, which were administered and scored under standardized testing conditions. This study is based on tasks developed for grades five and six by RAND and UCSB. All tasks in the experiment were developed in pairs based on general task descriptions called task shells. RAND developed two shells, inference and classification, and derived two 25-minute tasks from each shell. The two inference tasks were called Pendulum and Lever; the two classification tasks were called Animals and Materials. Brief descriptions of the RAND shells and tasks are contained in Appendix A. UCSB developed one shell that encompassed experimental design, observation, analysis, and application. They developed two tasks from this shell which were called Friction and Incline Plane. Each UCSB task had two 25-minute parts, which were administered sequentially. Therefore, a complete UCSB task required 50 minutes.

The conditions under which the tasks were developed, administered, and scored were quite similar to those that would be encountered in an operational testing program. In fact, the tasks whose costs are reported here were administered in conjunction with the regular administration of the 1992-93 CLAS statewide testing program.

Testing Chronology

The test development process followed the timeline described in Figure 1. Because we were working in conjunction with an operational state testing program, we had milestones to meet that were comparable to those that would affect developers of a large-scale assessment. The entire development, administration, and scoring process lasted slightly more than one year.

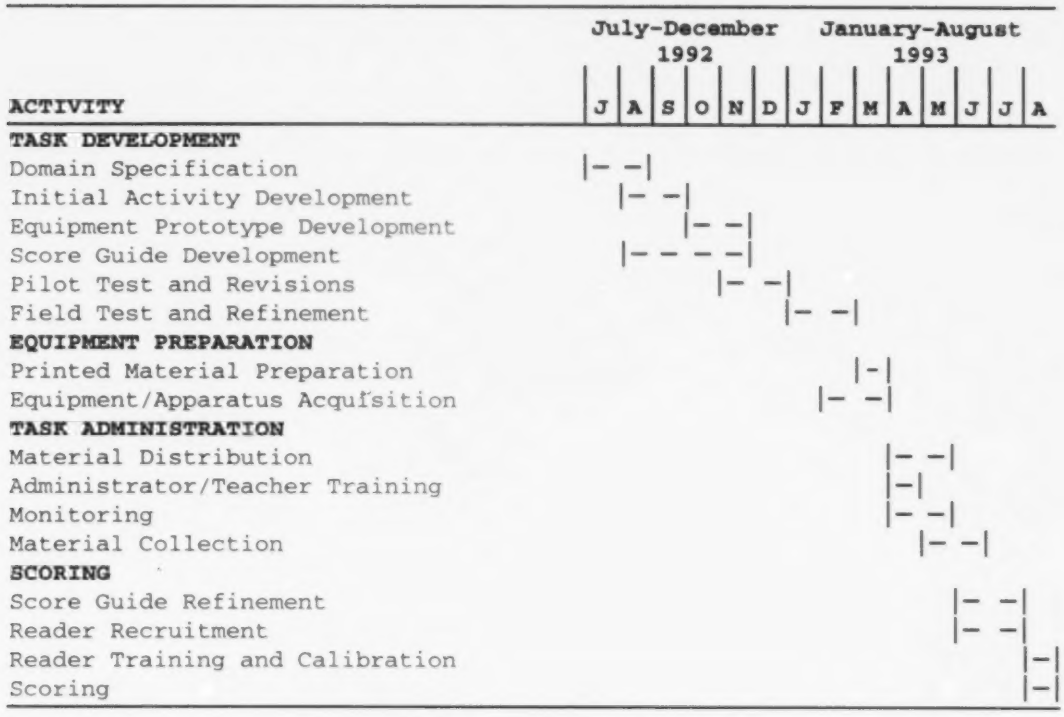

Figure 1-RAND Assessment Schedule

Sample

A sample of fifth and sixth grade classes from a diverse set of schools was selected for the study. (See Table 3.) A total of 2,200 students in two grade levels participated in the study.

Table 3

Sample for Science Performance Assessment Experiment

Unit	Grade 5	Grade 6
Schools	16	16
Classes	38	38
Students	1,100	1,100

Administration

As part of the experiment, each student completed five class periods of testing on five consecutive school days.[3] Fifth grade

[3]In most cases, assessments were administered at the same time each day, Monday through Friday. In very few cases, a weekend intervened between the third and fourth or fourth and fifth days.

students completed the two UCSB tasks, each of which required a full class period to complete, two days of CLAS testing (hands-on tasks and multiple choice questions) and the ITBS science test. Sixth grade students completed the four RAND tasks, which were administered in pairs, e.g., Lever followed by Animals, Materials followed by Pendulum, etc. in a counterbalanced design, two days of CLAS testing (hands-on tasks and multiple choice questions), and the ITBS science test. The CLAS and ITBS assessments were the same for both groups. Only the RAND and UCSB tasks are examined in our cost analysis.

All tasks were administered by Exercise Administrators (EAs) who were recruited, trained, and monitored by RAND. Testing materials were maintained at RAND and distributed to EAs on a weekly basis. The EAs transported materials by car to the testing sites and returned them to RAND at the end of the testing period. The EAs worked in two-person teams because of the large quantity of testing equipment and the limited school time available for set up prior to each test session and clean up thereafter.

Scoring

Student responses to the RAND and UCSB tasks were scored in the summer of 1993 in a single week-long scoring session. Readers were recruited, trained, and supervised by RAND staff. They received between one-half and one day of initial training and calibration. Following this initial training they scored student papers for three-and-one-half to four days. Supervisors monitored the scoring process, reading behind to determine if readers maintained standards, and holding recalibration sessions each day to insure the comparability of scores. On a typical day, readers devoted about an hour to calibration and about six hours to scoring. There were six groups of readers, one for each of the four RAND tasks and the two UCSB tasks. The number of readers in a group differed based on initial estimates of the time to score each task.

Other

Excluded from this analysis are all activities related specifically to the research component, such as experimental design, replacement of students names with identification numbers, and statistical analyses.

We did not budget for a number of other testing program functions that were unrelated to the experiment, including reporting scores, producing user manuals, and conducting public relations. Therefore, our estimates represent bare bones assessment costs.

RESOURCES AND COSTS

The final results of these analyses are presented in Table 4, and the derivations of these values are explained in the sections that follow. Table 4 shows the cost per student for the six hands-on science tasks based on three different size testing programs. The total cost for each task is estimated by combining task development, equipment preparation, task administration and scoring costs. The per student cost decreases as the size of the program increases. This pattern reflects assumptions about the relationship between costs and the number of students, which are described in the following sections.

Table 4

Per Student Cost of One Performance Task (Dollars)

					UCSB	
Number of	Inference		Classification		Friction	Incline Plane
Students	Pendulum	Lever	Materials	Animals	(I & II)	(I & II)
1,000	69	70	72	68	86	85
10,000	22	23	27	22	29	29
100,000	17	18	21	18	23	23

There are two ways to compute the average per student cost across tasks. The first approach is to compute the simple average of the six values, which ignores differences in the length of time required to complete each task. For a testing program serving 100,000 students the average cost per student per task is $20. An alternative value that we find more meaningful is the average cost of one period (approximately 45-50 minutes) of hands-on assessment. This value is computed by paring two RAND tasks from different shells and computing the mean based on these two pairs and the two UCSB tasks. Using this method (and assuming that 100,000 or more students are tested), the average cost per student

for one-period of hands-on science testing is $30 (or about 100 times more expensive than a multiple choice test of the same length).

The estimates of personnel resources are based on retrospective reconstructions of time allocations by RAND staff. Such accounts are imperfect, particularly since these activities took place over an extended period of time, and staff were engaged in other project activities (as well as other research) during this time. However, independent estimates from individual project staff were quite similar. Personnel resources for teachers and consultants as well as contracted costs and other direct charges are derived from actual expenditures.

Task Development

Task development costs are summarized in Table 5. The average task development cost was $45,636 per task or $68,453 per class period (combining two RAND tasks from different shells). As might be expected, the largest component of task development is personnel.[4] A five-person task development team worked on all the RAND tasks, with individual staff focusing their efforts on one or the other shell (inference or classification). The equipment costs in Table 5 reflect the development of the initial prototypes of the apparatus. Initial tasks were pilot tested in two schools; we included the teacher's time as a contributed resource (no resources were included for students' time). The other direct costs that appear in Table 5 for the RAND tasks were computing charges associated with the work of the RAND research staff.

The UCSB tasks were developed under subcontract to RAND, and the only information we have is the total costs of this effort. The UCSB tasks were twice the length of the RAND tasks, which explains, in part, the higher task development cost. (Detailed breakdowns of personnel

[4]Estimates of personnel costs are based on typical salaries for RAND staff at each grade level. External consultants were paid $125 per day, which included an allocation for providing a car and using it to transport testing materials from RAND to the school sites. Travel costs were estimated on the basis of mileage at the rate of $.26 per mile, and personnel costs were reduced by the amount allocated to travel. RAND overhead costs were added to all personnel and RAND fringe benefit costs were added to the RAND staff.

resources by staff level for each of the four major activities are presented in Appendix B.)

Table 5

Task Development Costs by Task (Dollars)

	Inference		Classification		UCSB	
Activity	**Pendulum**	**Lever**	**Animals**	**Materials**	**Friction (I & II)**	**Incline Plane (I & II)**
Personnel	41,136	41,136	38,882	38,882	4,361	4,361
Travel	0	0	0	0	0	0
Equipment	30	25	45	25	0	0
Contributed Time	63	63	63	63	0	0
Contracted Costs	0	0	0	0	50,000	50,000
Other Direct Costs	1,093	1,093	1,137	1,137	110	110
Total	42,322	42,317	40,126	40,106	54,471	54,471

Task development is expensive because extended hands-on tasks are complex, they must be designed to work within the constraints of the testing program, equipment must meet many standards, and student performance is extremely sensitive to subtle changes in format and presentation (Shavelson, et al., 1991). The complexity arises because these tasks reflect larger conceptual units than one normally associates with multiple choice items. For example, each UCSB tasks involved four components: planning and design, performance, analysis and interpretation, and application. It is extremely time-consuming to build a task that encompasses such a large domain. The testing environment places constraints on tasks that increase development costs. For example, tasks in a large-scale, on-demand, testing program must be completable in a fixed amount of time, all equipment must be portable, and activities must not require specialized facilities (e.g., sinks).

The use of equipment itself presents further problems. The apparatus must be tested to insure proper and safe operation. It must resist breakage and contamination, and the equipment must be available in large, standardized quantities. (Ask us about rocks that do not float, sandpaper that wears out, dropper bottles that leak, solutions that change properties while sitting on the shelf, strings that stretch, bugs that escape, and amphibians that change their color.) Finally,

every aspect of the presentation of the task affects students performance. The text, tables and diagrams must be reviewed and revised extensively to make sure they are interpreted correctly by students. We conducted structured "think-aloud" interviews with students who completed some of the RAND tasks and found considerable variation in students' understanding of the instructions (Hamilton, 1994).

Task development cost per student. The cost per student for task development can be computed by dividing the total cost by the number of students served. As the testing population increases, the per student cost of task development diminishes. (See Table 6.) With a large testing population, on the order of 100,000 students or more, the per pupil cost of task development approaches zero. As will be seen below, however, this decline in per student cost is not as dramatic for the other activities associated with hands-on testing.

Table 6

Task Development Cost Per Student for Three Testing Populations (Dollars)

					UCSB	
Number of Students	Inference		Classification		Friction (I & II)	Incline Plane (I & II)
	Pendulum	Lever	Materials	Animals		
1,000	42	42	40	40	54	54
10,000	4	4	4	4	5	5
100,000	<1	<1	<1	<1	1	1

Equipment Preparation

Equipment preparation costs are presented in Table 7. The total amounts represent the cost of producing kits of the RAND tasks for 80 students and UCSB kits for 90 students. Resources include staff time for coordinating, ordering, and assembling apparatus, and direct charges for equipment purchase and manufacture. (Resources for prototype development were included under task development.) In the case of the friction, incline plane, pendulum and lever tasks, we hired a local carpenter to construct multiple sets of equipment according to our specifications. In the case of the classification tasks, we located

commercial vendors for the objects, purchased them in bulk, and hired people to assemble the kits.

Table 7

Equipment Preparation Costs by Task for 80 Sets (Dollars)

	Inference		Classification		UCSB	
Activity	Pendulum	Lever	Animals	Materials	Friction (I & II)	Incline Plane (I & II)
Personnel	4,795	4,795	5,582	5,582	2,235	2,235
Travel	0	0	0	0	0	0
Equipment	750	920	1592	739	860	1,085
Contributed Time	0	0	0	0	0	0
Contracted Costs	0	0	0	0	0	0
Other Direct Costs	165	165	215	215	80	80
Total	5,710	5,880	7,389	6,536	3,175	3,400

*90 kits were prepared for each UCSB task.

Equipment preparation cost per student. Equipment preparation is divided into two components to compute the cost per student: the administrative resources needed for acquisition and coordination, and the equipment itself. We assume that acquisition costs are fixed and independent of the number of kits acquired, so these resources are divided by the number of students tested to compute the first component of per student cost. The second component, the cost of the apparatus itself, is computed by dividing the equipment cost by the number of kits produced to estimate the cost per kit, and dividing this figure by three to estimate the cost per student. (Since all the materials were reusable, the ratio of three uses per kit is used.) Finally, for large testing populations some economies of scale are factored into the estimates of the cost of apparatus. The cost of the equipment component is reduced 10% for 10,000 students and 20% for 100,000 students. (See Table 8.) Under these assumptions, the average cost per student per task for equipment for 100,000 students was $4; the average cost per student per class period (combining two RAND tasks from different shells or one UCSB task) was $6.

Table 8

Equipment Preparation Cost Per Student for Three Testing Populations (Dollars)

Number of Students	Inference		Classification		UCSB	
	Pendulum	Lever	Materials	Animals	Friction (I & II)	Incline Plane (I & II)
1,000	8	8	12	9	6	7
10,000	3	4	6	3	3	4
100,000	3	3	5	2	3	4

Task Administration

The four RAND and two UCSB tasks were administered as part of a five-day experiment that also included other hands-on science tasks and multiple choice science measures. On average, EA teams tested about 2.5 classes per day per school, and the same EA team administered all the assessment activities within a school. Resources for task administration were apportioned equally across the five days. The proportional share of administration costs for each task are presented in Table 9. The average total cost of administering a single task to 1,100 students was $13,888; the average total cost of for one class period of hands-on testing (combining two RAND tasks or using one UCSB task) for 1,100 students was $20,832.

Administration costs include recruiting and training Exercise Administrators (we contracted with a temporary agency for recruiting and initial screening of EA candidates), printing of testing materials, preparing and distributing materials, packing and transportation equipment, security dividers for test administration, travel to school sites, administering the tasks to students, monitoring task administration, and receiving complete materials and maintaining inventory. In those few cases in which travel distances were too great for daily commuting, we paid for overnight accommodations in the local area during the testing period.

Table 9

Task Administration Costs by Task for 1,100 Students (Dollars)

	Inference		Classification		UCSB	
Activity	Pendulum	Lever	Animals	Materials	Friction (I & II)	Incline Plane (I & II)
Personnel	11,311	11,311	11,311	11,311	15,405	15,405
Travel	215	215	215	215	429	429
Equipment	300	300	300	300	400	400
Contributed Time	0	0	0	0	0	0
Contracted Costs	225	225	225	225	450	450
Other Direct Costs	285	285	285	285	310	310
Total	12,335	12,335	12,335	12,335	16,994	16,994

The use of Exercise Administrators is uncommon in large-scale state assessments, and this approach to task administration may have added expenses to the experiment that would not be present in a state testing program. The costs and benefits of EAs are explored further in the discussion section.

Task administration cost per student. The per student cost of administration was computed by dividing the total cost by 1,100, the number of students in this experiment. We assume that these costs do not fluctuate based on the number of students. Therefore, the cost of task administration is $11 per student for each RAND task and $15 per student for the two UCSB tasks regardless of the number of students tested.

However, one component of administration, training, deserves special attention. We assumed that training costs would be directly proportional to the size of the testing population, i.e., there would be no economies of scale in training EAs beyond those we already realized from our group of 20 EAs. Even if this assumption is faulty and there are further economies to be achieved, training costs were a small fraction of total administration costs, so these economies will not affect the per student cost estimates in a substantial manner. More importantly, we suspect that as the number of EAs increases, more rather than less effort would be needed to insure standardization.

Scoring

All six tasks were scored during a week-long session in the summer of 1993. The major costs associated with scoring were personnel and facilities. Other costs include meals during the day long scoring sessions. Each student paper was read multiple times for research purposes. For this discussion, scoring costs have been adjusted to reflect the cost of reading each paper only once. The costs associated with a single reading of each task are summarized in Table 10. The differences in scoring costs in Table 10 reflect the differences in scoring time associated with each task.[5] The average cost per task for reading 1,100 papers one time was $8,306; the average cost per class period (based on two RAND tasks from different shells) was $12,459.

Table 10

Scoring Costs by Task for 1,100 Students (Dollars)

	Inference		Classification		UCSB	
Activity	Pendulum	Lever	Animals	Materials	Friction (I & II)	Incline Plane (I & II)
Personnel	7,332	7,148	7,684	6,876	8917	8143
Travel	78	78	150	103	166	103
Equipment	0	0	0	0	0	0
Contributed Time	0	0	0	0	0	0
Contracted Costs	0	0	0	0	0	0
Other Direct Costs	402	382	603	455	699	520
Total	7,812	7,608	8,437	7,434	9,782	8,766

The RAND approach to task development and scoring produced high reader reliability. Inter-reader correlations for all six tasks exceeded 0.90. These values suggest it is not necessary to read each paper more than once. If similar results can be obtained in an operational program, it would be necessary to re-read only a small fraction of the answers to verify reader accuracy. For example, double-scoring of 10% of the papers might be adequate for quality control

[5]On average, it required 1.8 minutes to score the inference tasks (lever or pendulum), 2.7 minutes to score the classifications tasks (animals or materials), and 3.1 minutes to score both parts (I and II) of the UCSB tasks.

purposes. Our cost analyses assume that 10% of the sample would be read a second time.

Scoring cost per student. The per student cost for scoring is computed by disaggregating scoring costs into two components and applying different procedures to each. (See Table 11.) The first component is preparation for scoring. Preparation includes those activities carried out by staff prior to the actual scoring session, such as reviewing scoring guides and sample papers, preparing calibration materials, estimating scoring times, and allocating readers to tasks. For research purposes, we constructed random batches of student papers and developed a scoring procedure for assigning batches to scorers. This procedure reduced spurious intra-class correlations among scores, and it is recommended for large-scale testing programs as well. However, we did not include the cost of these research activities in our estimates. In computing per student scoring costs, we assume that preparation costs are fixed regardless of the number of papers to be scored. Preparation costs are divided by the number of students to be tested to compute the first per student cost component.

The second component of per student scoring costs are the resources associated with the actual scoring sessions. These include time for readers and supervisors (e.g., table leaders) as well as facilities and meals. The total cost of providing one reading per paper was divided by the number of papers (1,100) to estimate the cost per student for reading and assigning a score. The two components were combined to estimate total scoring costs for the three different student populations.

Table 11

Scoring Cost Per Student for Three Testing Populations (Dollars)

Number of Students	Inference		Classification		UCSB	
	Pendulum	Lever	Materials	Animals	Friction (I & II)	Incline Plane (I & II)
1,000	8	8	8	7	10	9
10,000	4	4	5	4	5	4
100,000	3	3	4	3	5	3

CONCLUSIONS

Our experience suggests that hands-on science testing is about 100 times as expensive as standardized multiple choice testing for a comparable amount of testing time, and it is five to six times as expensive as constructed response assessments in writing of comparable length. Moreover, the economies of scale associated with performance testing are very limited. Although the task development cost per student drops dramatically as the number of students increases, the other major costs do not. Before considering some of the implications of these figures for large-scale assessment, it is important to review the assumptions under which they were made and to test the sensitivity of the results to changes in those assumptions.

Assumptions

These cost estimates depend on certain assumptions about staffing, task development, equipment, and task administration that warrant further discussion. Some might think that cost could be saved by reducing the amount or expertise of staff. Certainly, as a research project we were not subject to the same fiscal constraints that govern an operational testing program, nor did we feel the market pressures to reduce costs that affect the behaviors of commercial test publishers. As a result, it might be argued that the personnel estimates are inflated, e.g., the staffing levels were too high or the staff were overqualified. We would disagree on both counts. The principal investigator and the other senior staff member had extensive experience developing tests for large-scale testing programs, including constructed response items based on simulated performance situations. In fact, this expertise helped to reduce the resources required to develop the science tasks by laying out a clear, efficient process for task development similar to the methods used by large testing organizations. The fact that it was more difficult than anticipated to develop these hands-on tasks is a lesson well learned. Almost every aspect of the process, including task development, equipment, administration and scoring, required more time than planned based on previous test development experience. As to staff qualifications, we would argue that it is

crucial to have highly qualified staff. The complexity of this process increases the value of such expertise.

However, it is valuable to examine the effects of reductions in staff time or staff charges. The vast majority of RAND staff time (60%) went to test development. Because these costs are amortized over the entire testing population, even a fifty percent reduction in staffing for test development would not have an appreciable effect on the per student cost for large testing programs. The same is true for equipment acquisition and scoring preparation, which represent ten percent and nine percent of the professional staff time, respectively. A 25% reduction in RAND staff for task administration would reduce this component somewhat, but RAND staff represented less than one-half of the personnel for administration. (Changes in the use of Exercise Administrators are explored below.) We do not believe that scoring supervision resources can be reduced without adversely affecting the quality--reliability and validity--of the scores obtained. Overall, per student costs cannot be reduced substantially by reductions in personnel.

An additional concern is that there are errors in the retrospective estimates of staff resources. Admittedly, staff did not keep time sheets by task, and they were involved in a variety of project-related activities, so the estimates contain some error. We believe that the reconstructed resource estimates are likely to be low rather than high. We continually underestimated the complexity of working with hands-on assessment and the amount of time required to develop and administer these types of tasks. We tried to err on the conservative side when computing staff resources. Excluded from these estimates entirely was a qualitative study of student responses to the tasks which helped us improve them (Hamilton, 1994), as well as all of the time of the project statistician and data analyst who were involved in designing and preparing for the scoring sessions.

Equipment cost estimates are based on the assumption that kits would be reused three times. This assumption is based on the experience of the California Learning Assessment System. In an operational program, on-demand testing usually is confined to a narrow window of

time, so it is necessary to test in many classrooms simultaneously. Therefore multiple setups are required. In subsequent experiments we developed other tasks that had disposable components (e.g., solutions, testing papers, etc.), and each student required his or her own kit. Obviously, materials costs were higher for these tasks. A different testing scenario could reduce materials costs somewhat, but the three-to-one ratio seems like a reasonable middle ground for policy purposes. If one assumes that each kit could be used six times, the per student equipment cost could be cut in half, and the overall testing cost (on a per period basis) would be reduced by about 20%.

We assumed that tasks development costs are constant, i.e., it requires as much time to develop the second task from a shell as the first, and as much time to develop the second shell as the first. We treat task development costs as if they were independent of the number of tasks to be developed and the number of students to be tested. Some would argue that there are efficiencies that come from experience and scale. This principle suggests that task development costs might decrease as the number of tasks increases. In the case of the classification tasks, there were marked efficiencies associated with developing the second task from the shell; it required approximately one-half the time as the first task. However, this was not true for the inference shell. The second task (lever) required more time to develop than the first task (pendulum) because it was a substitute for an alternative task (involving particles in lake water) that did not prove to be feasible.

In addition, even the efficiencies associated with the classification shell may not continue indefinitely. After developing two or three tasks from a shell it becomes increasingly more difficult to find other ways to operationalize that shell. This principle holds more generally, even when one is not operating with shells. It is relatively easy to develop one or two science performance tasks, but as the number of desired tasks increases it becomes more difficult to find appropriate activities. There may be a limited number of tasks that are appropriate for large scale testing in that they meet content and skills demands and also fit the needs of a testing program for

transportability, safety, resistance to decay, etc. Therefore, we think the constant cost assumption is reasonable.

Our approach to task administration involved the use of Exercise Administrators rather than classroom teachers. This choice was dictated by the demands of the research, and it could be argued that this is not a reasonable model on which to base a large-scale testing program, although NAEP also uses this strategy. The salaries of the Exercise Administrators represented between 32% and 47% of the total cost for administration, which suggests that costs could be reduced substantially by having teachers do the administration. Before examining the validity of this claim, it is useful to note that even if the EAs were eliminated entirely, the task administration costs would be reduced by less than one-half and the average per student cost for 100,000 students would be reduced by $5 per task or $7 per class period, an overall reduction of about 25%. This amount is not insubstantial, but would not reduce costs enough to make this type of assessment affordable to most districts, states, or national programs.

One way to reduce EA costs would be to increase their efficiency. In this experiment, EAs tested about 2.5 classes per day. In an operational testing program, it might be possible to increase the number of classes assessed by an EA team each day. A twenty percent increase, from 2.5 to three classes per day seems easily attainable with more efficient scheduling. This would reduce assessment costs by a comparable amount. A 60% increase to four classes per day might be possible under ideal conditions, but the ideal would be difficult to achieve. Because of the need to set up equipment, testing an average of four classes per day would require that each school provide a separate room for testing and be willing to rearrange the normal class schedule to accommodate the assessment. We doubt that it would be possible to realize this level of efficiency on a large scale.

Furthermore, the costs of the EAs must be weighed against the benefits that accrue from their use as well as the additional costs that would have been associated with the use of classroom teachers. Using EAs provided important consistency in test administration. EAs were trained to set up and administer the complex hands-on assessments

according to precise guidelines and to respond to questions and unexpected conditions in a consistent manner, a necessary condition for meaningful comparisons between scores. The wisdom of this approach is borne out by the experience of CLAS. In the past, the CLAS program has found considerable between-teacher variation in administration even after teachers were trained. The National Assessment of Educational Progress has used Exercise Administrators throughout its twenty-year history for similar reasons.

Eschewing EAs would not be cost-free. There would be additional costs associated with the use of classroom teachers instead of EAs. Past experience with pencil and paper standardized tests was not sufficient to prepare a person to administer hands-on tasks. The cost of training classroom teachers would be substantial, including release time for all teachers to attend a one day training workshop (the minimum to prepare someone to administer approximately two-to-four hours of complex tasks). Furthermore, the number of teachers requiring training is many times as large as the number of EAs. In a large-scale testing program one would have to create an infrastructure for training, including a hierarchy of training supervisors, trainers of trainers, etc. The costs associated with setting up and maintaining this training system would be substantial. In addition, this is not just a one-time expense because tasks would presumably change every year or so and thus teachers would have to be trained for each new task. On the other hand, there are staff development benefits that accrue to teachers from direct involvement in the administration of hands-on science tasks. The value of this inservice training would have to be considered as well.

Overall, the assumptions under which this analysis was conducted appear to be reasonable, at least to us. Even if we were to adopt all the measures to reduce resource demands suggested in this discussion the maximum reduction in per student cost would be 50%. The resulting "bare bones" assessment would still cost about $15 per class period--about 50 times more expensive than multiple choice tests or three times more expensive than performance assessments in writing.

Implications for Large Scale Testing

It is impossible to interpret this information for large-scale testing without considering the question of reliability, both the agreement among readers in the score(s) they assign to a student (reader reliability) and the consistency of the scores assigned to a student across tasks (score reliability). Reader reliability for these tasks was quite high. This was achieved by preparing and testing scoring activities in the task development process, and by using focused semi-analytic scoring procedures. Inter-reader correlations above 0.9 on all six tasks indicate that it is only necessary to read each paper once to assign it a reliable score.

On the other hand, several tasks are needed to produce a reliable "hands-on" score for an individual student. The average inter-task correlation between hands-on assessments in the 1993 RAND experiment was about 0.5.[6] Under these conditions, approximately four tasks (two-to-three class periods) would be needed to produce a student score with a reliability that was comparable to the ITBS science subtest (i.e., about 0.8), and about eight or nine tasks (four-to-five class periods) would be needed to produce a student score with a reliability of 0.9 (which could be achieved with an ITBS science test of about one period in length). Consequently, the cost of a reliable hands-on score for each student would be between four and five times as great as the per period estimates presented here, i.e., between $120 and $150 per student.

It is unlikely that any jurisdiction is ready to spend $100 or more per student for science performance testing, so it is important to note that costs can be reduced by changing the level of reliability that is desired, the nature of score reporting, and the level of aggregation. For example, results from a single hands-on tasks could be combined with multiple choice scores to produce a reliable student score for about $30 per student. Aggregate hands-on scores at the school or district level could be produced much more inexpensively by a combination of sampling students and matrix sampling tasks to students.

---

[6]This was true in grade five and grade six, across RAND, UCSB and CLAS hands-on tasks, comparing two tasks from the same shell as well as two tasks from different shells.

Two other factors should be considered when thinking about the implications of this analysis for large-scale testing. First, start-up costs can pose a significant hurdle to some jurisdictions. By focusing on per student costs, this discussion has overlooked the substantial initial investment that is required for test development and test administration infrastructure. It would not surprise us to learn that state testing programs have considered and rejected various types of performance assessment because they could not afford the initial costs. Second, this analysis excluded many important elements of an operational testing program. We did not include estimates of the costs associated with activities such as test security, data analysis, score reporting, documentation, or public relations. We are developing a cost model for large-scale testing that is more comprehensive, and examines overall costs under different scenarios regarding test development, level of score aggregation, reliability, and program size.

REFERENCES

Hamilton, L.S. (1994). An investigation of students' affective responses to alternative assessment formats. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Shavelson, R J., Baxter, G.P., and Pine, J. (1991). Performance assessments in science. Applied Measurement in Education, 4(4), 347-362.

Shavelson, R J., Carey, N.B., and Webb, N.M. (1990). Indicators of science achievement: Options for a powerful policy instrument. Phi Delta Kappan, 71(9), 692-697.

Wiggens, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70(10), 703-713.

APPENDIX A:
DESCRIPTION OF RAND SHELLS AND TASKS

TASK SHELLS

All the tasks in the RAND experiment were developed in pairs based on a general task description called a task shell. A task shell is a defining sentence or model that specifies key characteristics of an investigation in conceptual terms and can be used to generate a number of specific tasks. We developed one shell that focused on inference and another shell that focused on classification. Each shell was used to develop two tasks.

INFERENCE SHELL AND TASKS

Students are presented with a situation in which they can measure/observe three variables, one dependent (outcome) and two independent. Both independent variables are plausibly related to the outcome, but only one of them is correlated with the outcome variable. The student is shown how to make the necessary measurements for one condition and record the results. The student is told to complete the experiment (make the measurements under the other three possible conditions) to determine whether one or both of the predictors affects the outcome. After collecting the relevant data, the student is asked to use the information gained from the experiment to identify which variable(s) affect the outcome and then to predict the behavior of the outcome variable under different conditions of the two predictors.

We derived two tasks from this shell both from the domain of physics. The first involved a pendulum, the second, levers. In the pendulum task a student was given two strings of different lengths, a "light" and a "heavy" weight, a stopwatch, and a bar from which strings and weights could be suspended and swung. The strings had loops at each end so they could be attached to the bar and the weights could be attached to them. The student was instructed on how to suspend a weight from a string to create a pendulum and then how to measure the periodicity of the pendulum by counting the time it took to swing back

and forth twenty times. The student was instructed to repeat the experiment for all four combinations of string and weights. Then the student was asked which influenced the time more, the length of string or the weight at the end of the pendulum. The student also was asked to justify their answer based on the experiment. Finally, the student was shown another pendulum with an intermediate length string and an intermediate weight and was asked to predict the amount of time it would take to swing back and forth twenty times without actually testing it.

The lever task was similar; the student was asked to determine whether the number of washers required to lift a fixed weight was determined by the length of the lever and/or the proportion of the bar on the side of the fulcrum where the weights were placed. To conduct this experiment, the student was given four bars (two short and two long) with pivot notches either one-half or one-quarter of the way down the bar. After testing these four bars, the student was shown a fifth bar and asked to predict the number of washers required to lift the weight using this bar without testing it.

CLASSIFICATION SHELL AND TASKS

Students are taught about two-way cross-classification using objects. They perform a simple "tuning" task in which they are led through the development of a two-way classification system and they sort objects into four mutually exclusive groups. They are given a new set of objects that differ in a number of ways and asked to construct their own two-way classification system and sort the objects appropriately using any relevant features of the objects as the classification dimensions. The only restrictions are that each object had to be put in one of the four cells and each cell had to have at least one object in it. Finally, students are given an additional object that had been concealed, and they asked to classify it using their system or to explain why it does not fit.

Two tasks were developed from this shell. In the tasks, two-way classification was explained using pictures of people who differ in terms of age (old and young), position (sitting or standing), gender (male or female), and type of clothing (summer or winter). Students

were shown how to classify the pictures into four mutually-exclusive groups using two dimensions simultaneously (e.g., young-males, old-males, young-women, and old-women). This activity was the same in both tasks. In the "Animals" classification task, all the objects were plastic land and sea animals, and the extra animal was an amphibian. In the "Materials" task, all the objects were natural materials, including rock, bone, animal hair, shells, sand, etc.; and the extra object was a manufactured pencil.

APPENDIX B:
DETAILS OF PERSONNEL RESOURCES

The RAND project staff consisted of two senior researchers, two junior researchers, one research assistant, one senior statistician, and three secretaries. We hired substitute teachers and other individuals with prior classroom experience to serve as Exercise Administrators during the testing phase. We also used teacher consultants during the test development phase and during scoring.

The following tables contain detailed information about the amount of time each class of individuals devoted to each of the four activities, task development (Table B.1), equipment preparation (Table B.2), task administration (Table B.3) and scoring (Table B.4).

Table B.1

RAND Personnel for Task Development (Person Days)

	Inference		Classification		UCSB	
Personnel	Pendulum	Lever	Animals	Materials	Friction (I & II)	Incline Plane (I & II)
Senior Staff	33	33	27	27	3	3
Junior Staff	7	7	11	11	3	3
Research Assistant	5	5	10	10		
Clerical	10	10	10	10		
Teacher Consultants	2	2	2	2		
Total	57	57	60	60	6	6

NOTE: RAND staff made only minor change on the UCSB tasks.

Table B.2

RAND Personnel for Equipment Preparation for 80 Kits (Person Days)

	Inference		Classification		UCSB	
Personnel	Pendulum	Lever	Animals	Materials	Friction (I & II)	Incline Plane (I & II)
Senior Staff	2	2	1	1		
Junior Staff	2	2	3	3	1	1
Research Assistant	2	2	4	4		
Clerical	3	3	3	3	3	3
Teacher Consultants	1	1	1	1	1	1
Total	10	10	12	12	5	5

NOTE: RAND staff made only minor changes on the UCSB equipment.

Table B.3

RAND Personnel for Task Administration for 1,100 Students (Person Days)

	Inference		Classification		UCSB	
Personnel	Pendulum	Lever	Animals	Materials	Friction (I & II)	Incline Plane (I & II)
Senior Staff	3	3	3	3	3	3
Junior Staff	5	5	5	5	6	6
Research Assistant	4	4	4	4	4	4
Clerical	3	3	3	3	3	3
Teacher Consultants	16	16	16	16	32	32
Total	31	31	31	31	48	48

Table B.4

RAND Personnel for Scoring for 1,100 Students, Adjusted (Person Days)*

	Inference		Classification		UCSB	
Personnel	**Pendulum**	**Lever**	**Animals**	**Materials**	**Friction (I & II)**	**Incline Plane (I & II)**
Senior Staff	3	3	1	1	1	1
Junior Staff	1	1	1	1	4	4
Research Assistant	1	1	1	1	1	1
Clerical						
Teacher Consultants	6	6	12	8	13	8
Total	11	11	15	11	19	14

NOTE: *Adjusted to reflect one score per student with 10% rescoring